**Supplementary material to Whitney, K. D., B. Boussau, E. J. Baack, and T. Garland Jr. *in press*. Drift and genome complexity revisited. *PLoS Genetics*.**

Tree topologies

Two topologies were examined, one favoring the Coelomata hypothesis (Figs. S1a and S2a), in which protostomes are not monophyletic, and one favoring the Ecdysozoa hypothesis (Figs. S1b and S2b), in which protostomes are monophyletic (Adoutte et al. 2000).

Branch lengths

Fossil estimates -- Divergence dates (in billion years before present) were gathered from the literature (Brocks et al., 2005; Battistuzzi et al., 2004; Douzery et al., 2004; Benton and Donoghue, 2007) and from http://www.fossilrecord.net/, and were applied to the two tree topologies. Twenty-two of the 28 total nodes in the phylogeny were dated in this way. The remaining six nodes were placed evenly between neighboring dated nodes (see BLADJ procedure of Webb et al. 2008). The resulting branch lengths are shown in Fig. S1.

rRNA estimates -- Branch lengths based on expected numbers of substitutions in ribosomal RNA were determined and applied to the two topologies. Aligned 18s or 16s (small subunit) ribosomal RNA sequences were downloaded from the SILVA database (http://www.arb-silva.de/, Pruesse et al., 2007) for the 29 studied species. Sequences from *Trypanosoma vivax* and *Tetraodon nigroviridis* were used to represent respectively *Trypanosoma sps.* and *Fugu rubripes*, for which sequences could not be found in the database. Based on this alignment, branch lengths were estimated in the maximum likelihood framework using bppml (Dutheil and Boussau, 2008) using a GTR + G(4 categories) + Invariant model of sequence evolution. The resulting branch lengths are shown in Fig. S2.

Regression analyses (corresponding to Table 2 of main paper)

The relationship between $N_e u$ and genome size for n=29 taxa was examined using REGRESSIONv2.m (Lavin et al. 2008) running in MATLAB v. 7.9.0. Three types of models were examined: ordinary least squares (OLS), phylogenetic generalized least squares (PGLS), and phylogenetic regression in which the residual variation is modeled as an Ornstein-Uhlenbeck process (RegOU). OLS is traditional 'nonphylogenetic' regression, which in effect assumes a star phylogeny in which all species are equally unrelated, and corresponds to the $N_e u$ vs. genome size analysis reported in Lynch & Conery (2003). PGLS assumes that residual variation among species is correlated, with the correlation given by a Brownian-motion like process along the specified phylogenetic tree (topology and branch lengths). The RegOU model estimates (via restricted maximum likelihood) the strength of phylogenetic signal in the residual variation simultaneously with the regression coefficients; the former is given by d, the Ornstein-Uhlenbeck transformation parameter. An OU evolutionary model is typically used to model the effects of stabilizing selection around an optimum. When d=0, there is no phylogenetic signal in the residuals from the regression model; when d is significantly greater than 0, significant phylogenetic signal exists (Lavin et al. 2008; Blomberg et al. 2003).

In addition to the single OLS analysis, twelve analyses were done. These corresponded to PGLS and RegOU models each run on the six combinations of two topologies (above) and three sets of branch lengths (all=1, as in Whitney & Garland 2010; and Fossil and rRNA, as above).

We compared the likelihoods of the PGLS and OLS models, with a higher likelihood taken as evidence of a better-fitting model. OU and OLS models were compared with ln maximum likelihood ratio tests and $P<0.05$ as a cutoff.

Regression analyses for bacteria

The relationship between $N_e u$ and genome size for n=7 bacterial species from the dataset of Lynch and Conery (2003) was examined using OLS and PGLS in REGRESSIONv2.m (Lavin et al. 2008) as described above using all=1 branch lengths.

$N_e u$ threshold values

Observations from the dataset of Lynch & Conery (2003) were scored for "low" vs. "high" $N_e u$ according to the following thresholds identified in Lynch & Conery (2003): intron number $N_e u = 0.015$; total transposon number (and fraction) $N_e u = 0.0128$. While the former was presented in the text, the latter was determined from Fig. 4 (bottom right panel) of Lynch & Conery (2003) using the program g3data 1.5.1 (http://www.frantz.fi/software/g3data.php) to estimate genome size at the threshold as 11.89 Mb. $N_e u$ was then back-calculated from genome size using the regression equation in the Fig.1b legend of Lynch & Conery (2003).

**References**


Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, de Rosa R. 2000. The new animal phylogeny: Reliability and implications. Proceedings of the National Academy of Sciences of the United States of America 97(9): 4453-4456.

Battistuzzi FU, Feijao A, Hedges SB. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evol Biol. 4:44. (2004).

Benton MJ, Donoghue PC. Paleontological evidence to date the tree of life. Mol Biol Evol. 24(1):26-53. (2007)

Blomberg SP, T Garland, and AR Ives. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution 57:717-745.

Brocks, Jochen J, Love, Gordon D, Summons, Roger E, Knoll, Andrew H, Logan, Graham A, et Bowden, Stephen A. Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. Nature, 437(7060), 866-870. (2005).

Douzery EJ, Snell EA, Bapteste E, Delsuc F, Philippe H. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils?
Proc Natl Acad Sci U S A. 101(43):15386-91. (2004).

Dutheil J, Boussau B. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evol Biol. 8:255. (2008).

Lavin SR, Karasov WH, Ives AR, Middleton KM, Garland T (2008) Morphometrics of the avian small intestine compared with that of nonflying mammals: A phylogenetic approach. Physiological and Biochemical Zoology 81: 526-550.

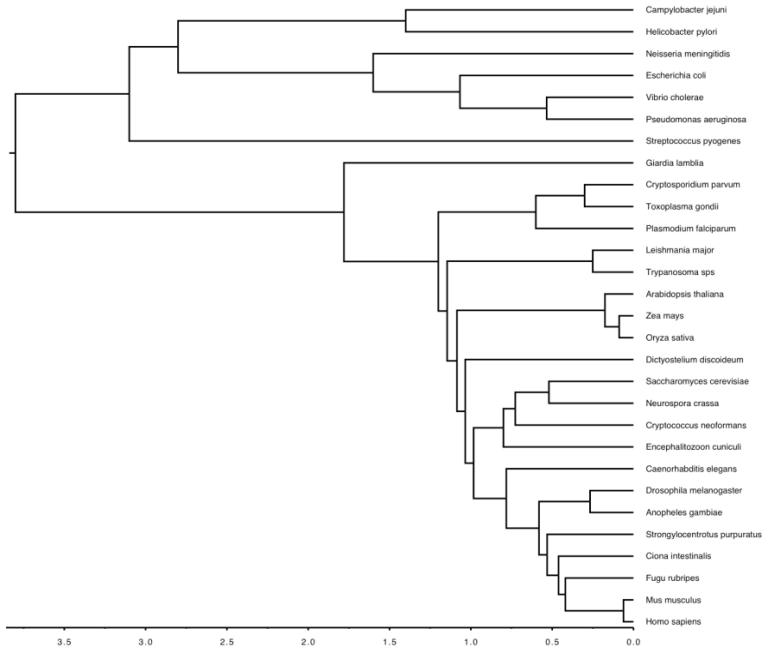
Lynch M, Conery JS (2003) The origins of genome complexity. Science 302: 1401-1404.

Pruesse, E., C. Quast, K. Knittel, B. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nuc. Acids Res. 35(21):7188-96. (2007).

Webb CO, Ackerly DD, Kembel SW (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics 24: 2098-2100.

Whitney KD, Garland Jr. T (2010) Did genetic drift drive increases in genome complexity? PLoS Genetics 6: e1001080.

a)




b)

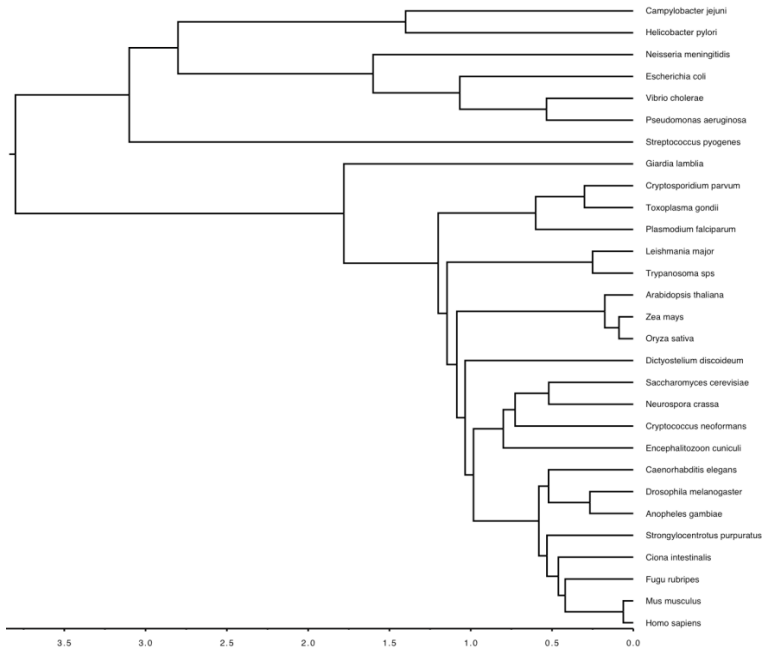
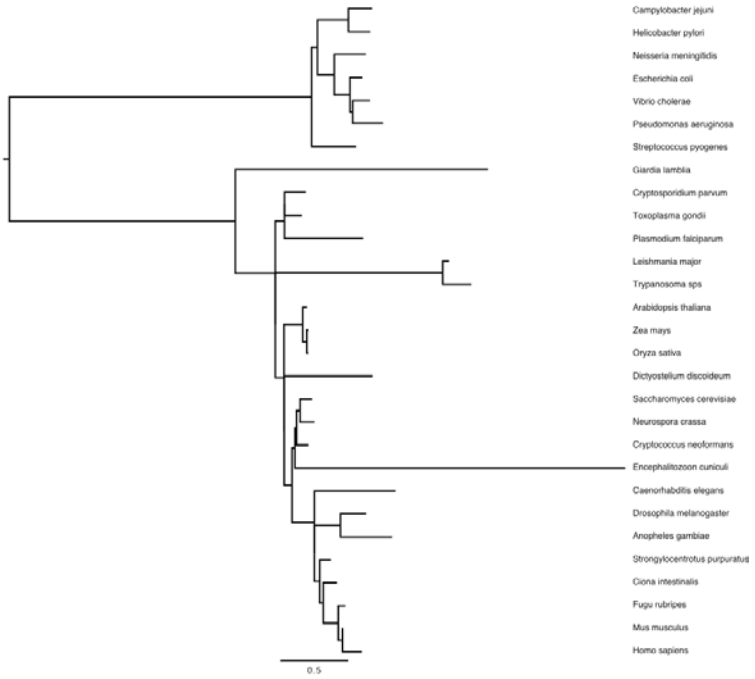



**Figure S1: Branch lengths based on the fossil record.** Dates are in billion years. a) Tree topology according to the Coelomata hypothesis. b) Tree topology according to the Ecdysozoa hypothesis.

a)




b)

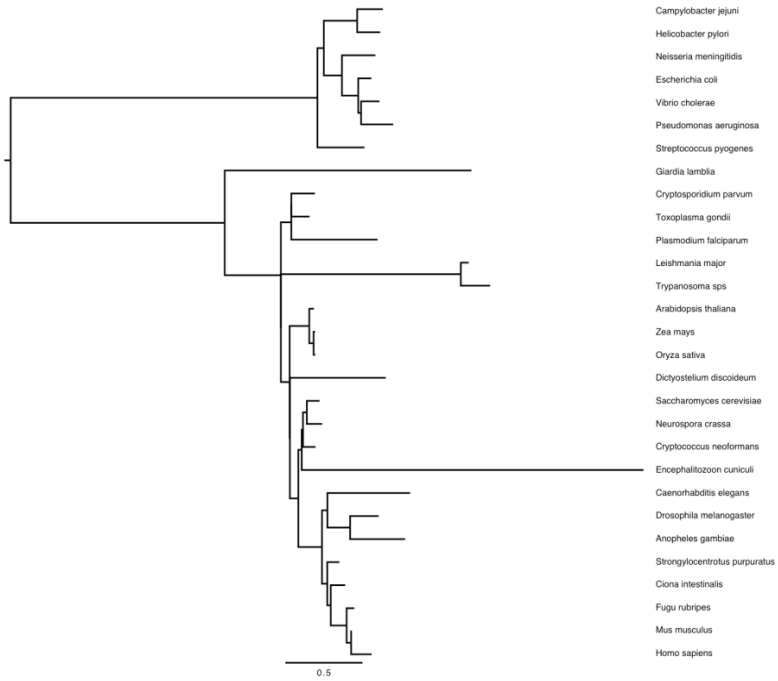



**Figure S2: Branch lengths based on expected rRNA substitutions**. a) Tree topology according to the Coelomata hypothesis. b) Tree topology according to the Ecdysozoa hypothesis.